Appendix A

Explanation of Item Response Theory Concepts Used in Analyzing Potential Test Score Accuracy

Explanation of Item Response Theory Concepts Used in Analyzing Potential Test Score Accuracy

The analyses of potential test score accuracy reported in Chapter 4 rely heavily on item response theory (IRT) models. Three constructs are important to understanding our analyses: item characteristic curves, test characteristic curves, and conditional standard errors. What follows is an attempt to describe these concepts without getting bogged down in the underlying assumptions and mathematical formulae.

Item Characteristic Curves. Item characteristic curves are functions that predict the probability of passing a given item for all students at a given level of ability. In IRT models, this probability depends on the *relative* difference between the item's difficulty and the examinee's ability. Initially, neither item difficulty nor examinee ability are known, so ability is placed on an arbitrary scale and item difficulties are "scaled" (estimated) relative to these abilities. Most commonly, the examinee ability is assumed to have a normal ("bell-shaped") distribution with a mean (average score) of zero and a standard deviation of 1.0. Figure A.1 shows an example of this normal distribution, where the height of the normal curve indicates the relative frequency of students with that level of ability.

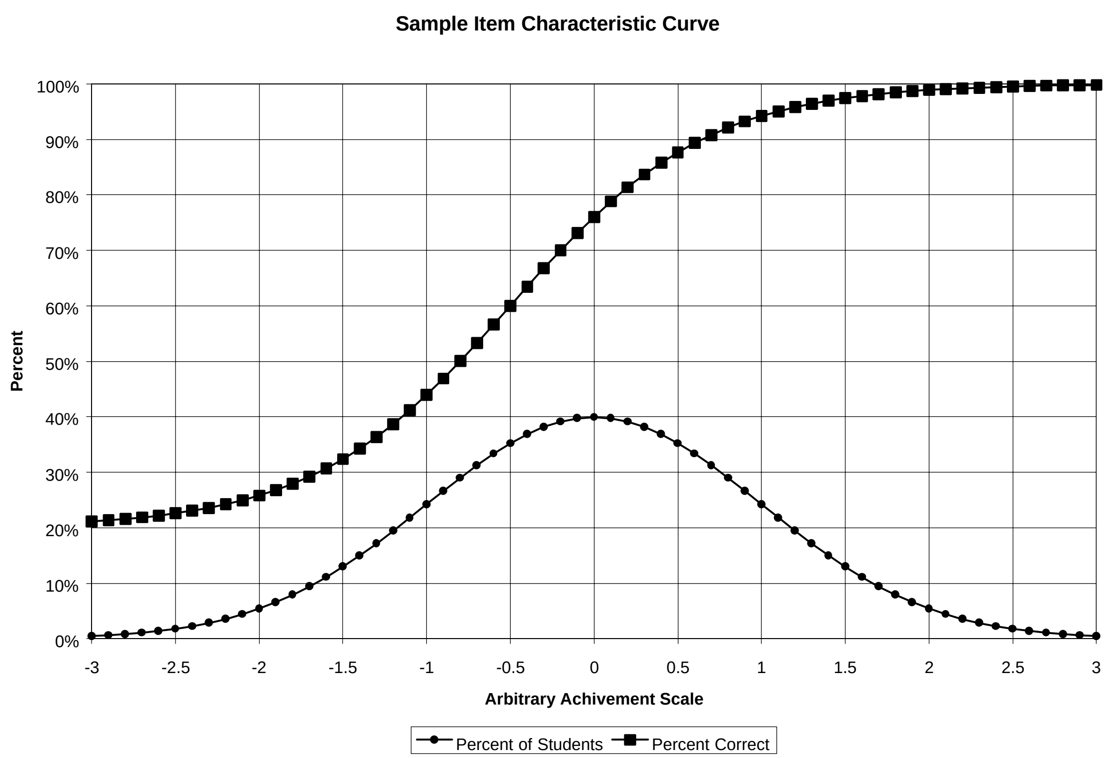

Figure A.1. Illustration of IRT concepts of examinee ability distribution and the probability of passing an item.

Figure A.1 also shows an example of the most common form (the three-parameter-logistic or 3PL model) of an item characteristic curve (ICC). This function has three parameters that capture important differences in the relationship of ability to the probability of correct responses for different items. ICCs are generally S-shaped curves that vary between a lower asymptote and 1.0. The lower asymptote (or "c" parameter) is the probability of answering correctly for students with no ability at all. For multiple choice items, the lower asymptote is generally close to one divided by the number of possible response options. This is the probability of getting the correct answer through random guessing. The c-parameter will vary also across items as a function of the allure or repugnancy of the different incorrect choices. The item in Figure A.1 has a c-parameter of .20.

The other two parameters that define the shape of different ICCs are the difficulty ("b") and slope ("a") parameters. The b-parameter is the point on the ability scale at which the probability of a correct response is half way between the lower asymptote and 1.0. As the b-parameter increases, the ICC moves to the right and the probability of a correct response at each ability level goes down, consistent with the general ideal that passing rates are lower for more difficult items. The item in Figure A.1 has a b-parameter of –0.5. The slope ("a") parameter gives the slope of the curve at the ability given by the b-parameter. It is sometimes call a discrimination parameter because it indicates how well the item discriminates between high and low ability students. (This has nothing to do with discrimination among students based on other characteristics.) For items with high a-parameters, the probability of a correct response drops off (increases) sharply for students below (above) the ability level defined by the b-parameter.

Test Characteristic Curves (TCC). Test characteristic curves are similar to item characteristic curves except that what is predicted is the total number of items in a whole test that an examinee at a given ability level will answer correctly rather than the probability of answering an individual item correctly. The expected number correct is simply the sum of the probabilities of answering each of the individual items correctly, so TCCs are computed by summing the ICCs. For example, for a three-item test, a given examinee's probability of answering the first item correctly was .25, the probability for the second item was .50 and the probability for the third item was again .25. This examinee's expected or average score would be 1.0 (.25 + .5 + .25). Note that the estimated scores are averages and do not have to be whole numbers.

Conditional Standard Errors. In the preceding example, an examinee of a given ability had an expected score of 1.0 on a three-item test. This does not mean that the examinee will get exactly one item correct every time. Some of the time (about 28%) he or she will miss all of the items and some of the time (about 3%) he or she will answer all of the items correctly. The "standard error of measurement" provides an indication of how often an examinee's actual score on a particular test form will differ significantly from their expected score (across all parallel forms).

For normative tests, it is common to summarize measurement error by a single number. The *test reliability* of a test is defined for a specific population equivalently as the expected correlation of scores from parallel forms or the ratio of the variance of the true scores to the total (true and error) score variance. From this last definition, the reliability is said to give the proportion of total variance "accounted" for by the underlying ability. (Note that the proportion not accounted for is error.) Coefficient alpha (Cronbach, 1951) is a statistic used to estimate test reliability by summarizing the agreement among a set of items.

For tests such as the HSEE, which are used to make important decisions at particular levels of achievement, overall test reliability is not the main issue. The critical concern is accuracy at the specific point on the score scale where a pass/fail decision must be made. The idea of conditional standard errors, is that error of measurement can be estimated separately for ("conditioned on") each level of ability.

IRT provides a basis for estimating conditional standard errors. Because of independence assumptions, it is possible to compute the probability of each possible number correct score for a student at a given achievement level (once the item parameters have been estimated). From this information, it is possible to determine how often observed scores will vary from the true score by any given amount. The conditional standard error is defined such that the absolute difference between the observed and true scores will be less than one standard error about two-thirds of the time. If errors are distributed normally, the difference will be less than two standard errors about 95% of the time and this criterion is used most often in constructing confidence bands.